УДК 003.26+347.78

Н. П. Шутько, аспирант (БГТУ)

ОСОБЕННОСТИ И ФОРМАЛЬНОЕ ОПИСАНИЕ ПРОЦЕССА ОСАЖДЕНИЯ СЕКРЕТНОЙ ИНФОРМАЦИИ В ТЕКСТОВЫЕ ДОКУМЕНТЫ НА ОСНОВЕ СТЕГАНОГРАФИИ

Объект исследования данной статьи — текстовые документы, обрабатываемые с помощью известных текстовых процессоров и редакторов, а также коды компьютерных программ и файлы баз данных. Описаны основы математического моделирования процесса осаждения тайной авторской информации в указанные типы документов на основе текстовой стеганографии. Тайная информация предназначена для защиты прав интеллектуальной собственности. Осаждение информации предусматривает изменение цветовых координат символов текста. Основу математической модели составляют пространственные координаты и цветовые параметры пикселей, формирующих растр текста.

The object of the research of this article — text documents that are processed with the help of some word processors and editors, as well as codes of software and database files. The basics of mathematical modeling of the embedding process of the secret author information in these types of documents based on text steganography are described. Secret information is intended to protect intellectual property rights. An Embedding provides changing the color coordinates of characters in the text. The base of mathematical model is the spatial coordinates and color values of pixels forming the raster of the text.

Введение. В настоящее время остро стоит проблема охраны авторского права. Как известно, интеллектуальная собственность — право на результат собственного интеллектуального труда, включая литературный труд, объекты науки, электронные документы и т. д.

Документом, регулирующим взаимоотношения между автором и издателем, является закон Республики Беларусь от 17 мая 2011 г. № 262-З «Об авторском праве и смежных правах».

Авторское право распространяется на произведения науки, литературы и искусства, находящиеся в какой-либо объективной форме:

письменной (рукопись, машинопись, нотная запись);

электронной (компьютерная программа, электронная база данных, текст);

звуко- или видеозаписи (магнитная, оптическая, электронная);

изображения (картина, рисунок, кино-, теле-, видео-, фотокадр);

объемно-пространственной (скульптура, макет, сооружение).

Охрана авторского права на печатное издание регламентируется авторским договором. Издание (переиздание) произведения должно осуществляться с согласия автора. Однако известны многочисленные случаи отступления от этого требования. Тем более, что довольно проблематично отследить использование того или иного электронного документа, его тиражирование. Система защиты авторских прав в Интернете в Беларуси еще на стадии становления.

В контексте рассматриваемой проблемы следует отметить еще одно важное обстоятельство. Цифровые технологии преобразования бумажных документов в электронную форму используются не только в офисном документообороте. Многие учреждения и частные лица сканируют бумажные документы, после чего последние могут храниться, архивироваться и более легко находиться. За последние несколько десятков лет принтеры и ксероксы стали общедоступными, простыми в употреблении и обслуживании и позволяют с помощью соответствующих программных средств производить разнообразные манипуляции с цветовыми параметрами текстов или графики. Это, по существу, нивелирует различие между бумажным и электронным документами и означает, что документы на бумажных носителях могут сканироваться и пересылаться по сетям связи почти также легко, как и изначально созданные электронные документы.

Таким образом, на данный момент угроза информационного пиратства в отношении электронных и бумажных документов приобрела практически одинаковую остроту. Эта особенность связана с возможностями современных компьютерных (цифровых) технологий.

В последнее время активно разрабатываются и исследуются методы обеспечения прав интеллектуальной собственности на основе стеганографии [1, 2]. Нами предложен ряд стеганографических методов, предназначенных для защиты электронных документов [3–6].

Для сравнительной оценки эффективности того или иного разработанного метода по критерию максимального объема осаждаемой информации, или стегостойкости (аналог понятия «криптостойкость»), необходим соответствующий математический аппарат. Общая идея, которую мы используем для решения указанной задачи, основана на формальном описании процедуры осаждения цифрового водяного знака (ЦВЗ) [7].

Далее в статье будут описаны важные, с нашей точки зрения, аспекты, относящиеся к разработке математической модели процесса стеганографической защиты текстовых документов.

Основная часть. Представляем набор приемов для встраивания неразличимых знаков (авторской информации) в отформатированные документы. Формально указанный прием связан с некоторым преобразованием или кодированием символа текста. Или более точно: в нашем случае кодирование символа — это изменение детальной особенности этого символа. Примеры возможных изменений детали включают изменение высоты отдельного знака или его позиции относительно других символов (например, настройка кернинга) или изменение параметров, формирующих цвет символа. Понятно, что для получения идентичной авторской информации в процессе ее извлечения аналогичные параметры какой-то части символов текста должны быть неизменными. Эти символы, таким образом, являются как бы «реперными точками», позволяющими производить сравнительные операции.

Фундаментальной особенностью шрифтов является отделение информации о форме символов от процесса их воспроизведения на растровом выводном устройстве. Если контуры символов шрифта можно описывать самыми разными способами, то задача воспроизведения, в конечном итоге, сводится к активизации некоторых точек (высвечиванию пикселей на экране дисплея или заполнению краской при печати на принтере).

Нетрудно заметить, что при обратном преобразовании (извлечении сообщения) могут возникнуть неоднозначности, обусловленные структурой (гарнитурой) шрифта [8].

Такая неоднозначность может быть вызвана, например, нарушением симметричности некоторых символов текста (к примеру, возникновением разного расстояния между вертикальными штрихами буквы Ш), что резко искажает их форму и затрудняет восстановление авторской информации.

Еще одной особенностью является то, что при воспроизведении символов на устройствах с малой разрешающей способностью (до тысячи пикселей на дюйм), особенно при выводе текста небольшим кеглем (12 и меньше), сильно сказываются ошибки масштабирования. Масштабирование происходит в абсолютных координатах относительно некоторой произвольной точки и всегда приводит к получению целочисленного результата. При этом возникает проблема округления нецелых результатов. Например, если координаты некоторого элемента символа в системе координат описания контура равны (200; 100), то при уменьшении размера контура в 3 раза они трансформируются в (66,66; 33,33). Поскольку нам нужны целые значения, они превратятся в (67; 33), т. е. значение горизонтальной координаты немного (на треть пикселя) увеличится, а вертикальной — на столько же уменьшится.

Существуют и другие особенности (выпадение точек, нарушение формы округлых букв — В, О, Р, С, б, е и др.). Однако это касается, в первую очередь, таких методов текстовой стеганографии, как *Line Shift Coding* и *Word Shift Coding* [3].

Для дальнейшего анализа в качестве основного элемента символа текста, с помощью которого мы осуществляем упомянутое кодирование, определяется цвет.

Текстовые документы-контейнеры, в которые осаждается тайная (авторская) информация, создаются, например, в текстовом процессоре MS Word. Будем считать, что исходный или стандартный цвет символов во всем текстовом документе — черный, и отождествлять текст с графическим рисунком.

Графические цветные файлы со схемой смешения RGB кодируют каждую точку рисунка тремя байтами. Данная точка (положим, это пиксель) состоит из трех составляющих: красного, зеленого, синего. Изменение каждого из трех наименее значащих бит (по известному методу LSB) приводит к изменению менее 1% интенсивности точки. Это позволяет скрывать в стандартной графической картинке объемом 800 кбайт около 100 кбайт информации, что не заметно при просмотре текста.

Вся информация, выводимая на экран компьютера, имеет двоичный вид, т. е. представляет собой совокупность «0» и «1». Для примера закодируем «0» и «1» различными цветами. Пусть мы хотим встроить в электронный текстовый документ (первая страница данной статьи) сообщение «Труды БГТУ». Данное сообщение в двоичном коде будет иметь вид, показанный на рис. 1.

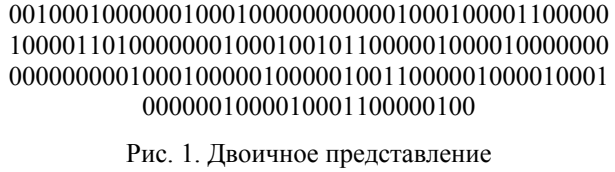
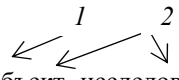


Рис. 1. Двоичное представление осаждаемого сообщения

Для наглядности процесса встраивания пусть «0» будет иметь цветовые координаты RGB (0; 100; 0), а «1» — (100; 0; 0).

Скрытие данных производится не только в обычных, но и в специальных (мягкий перенос, разрыв строки и др.) символах и пробелах.

Результат встраивания стегосообщения в текстовый документ на основе кодирования цветового параметра соответствующего символа показан на рис. 2. Процесс стегопреобразований выполняется с помощью специального программного средства [9].



Объект исследования данной статьи — текстовые документы, обрабатываемые с помощью известных текстовых процессоров и редакторов, а также коды компьютерных программ и файлы баз данных.

Рис. 2. Текстовый документ
с осажденной информацией

Символы маркируются изменением их цветовых координат в большую или меньшую сторону. Соседние символы не маркируются. На последнем рисунке выбор кодируемых символов производился псевдослучайным образом, что составляет один из элементов ключа.

На рисунке некоторые буквы выделились красным (соответствует «1») и зеленым (соответствует «0») цветами. В черно-белом формате данной статьи обнаружить различие цветовых параметров (красный или зеленый) символов достаточно сложно. Поэтому обозначим их: *1* — указывает на символ красного цвета, а *2* — зеленого.

В последующем, чтобы извлечь стегосообщение, необходимо анализировать информацию о цветовых координатах символов в документе, и, таким образом, с помощью известного только автору секретного ключа восстанавливается тайное сообщение.

Существующий подход формального представления процесса прямого и обратного стегопреобразований текстового документа, известный, например, из работы [10], предусматривает анализ так называемых горизонтальных и вертикальных профилей выделенного фрагмента текста. Эти профили основаны на подсчете пикселей в каждом (x) горизонтальном ($x = 1, 2, \ldots, L$; L — количество «черных» пикселей в одном горизонтальном ряду черно-белого растра, формирующего анализируемый фрагмент текста) или в каждом (y) вертикальном ряду ($y = 1, 2, \ldots, W$; W — количество «черных» пикселей в одном вертикальном ряду такого растра). В сравнении с простой такая модель не позволяет напрямую учитывать, например, цветовые параметры символов.

Дальнейшие наши рассуждения будем строить, исходя из нескольких понятных положений:

1) используется трехканальная (RGB) модель формирования растра (bitmap), соответствующего текстовому документу;

2) анализируемый фрагмент текста представляется в виде матрицы ($W \times L$) пикселей, местоположение каждого из которых определяется параметром $p(x, y)$;

3) влияние упомянутых выше особенностей шрифтов (нарушение симметричности, недостатки масштабирования, выпадение точек, нарушение формы округлых букв и др.) на процесс стегопреобразований текста будем рассматривать как наложение шумовой компоненты $N(x, y)_{rk}$ на соответствующие пиксели, составляющие контуры данного символа (находится в r-й строке и k-м столбце текста);

4) общая структурная схема стеганографической системы соответствует рис. 3.

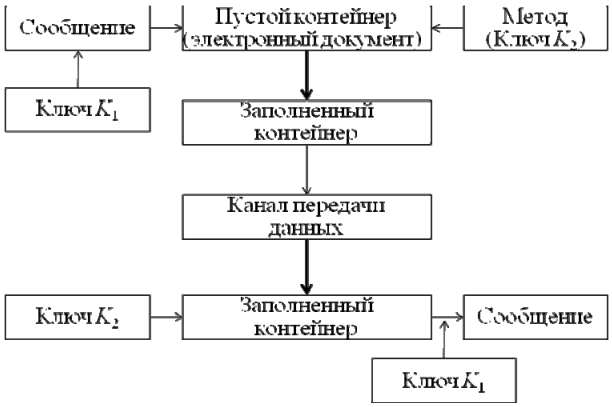


Рис. 3. Структурная схема
стеганографической системы

Осаждаемая авторская информация (сообщение) шифруется с использованием какого-либо ключа K_1. Полученное сообщение осаждается в контейнер с применением ключа K_2. Такой ключ может быть многопараметрическим (определяется используемым методом стегопреобразования или, например, выбором кодируемых символов на основе псевдослучайности и др.).

Будем считать, что заполненный контейнер не претерпевает изменений в ходе передачи данных. Заключительный этап — извлечение стегосообщения и его дешифрование.

За основу формального описания процесса нами взята методология описания процессов генерации и осаждения ЦВЗ в электронные документы, а также извлечения из них.

Пусть компонентами формального описания являются:

– множество скрываемых (авторских) сообщений M;

– множество возможных контейнеров (текстовых документов) T;

– множество возможных ключей (методов генерации авторской информации, например зашифрованной) K_1;

– множество возможных ключей (методов осаждения авторской информации) K_2.

Тогда формально процессы можно описать следующим образом:

– создание стегосообщения M_S:

$$M_S = F\left(M,\ K_1\right);$$

– создание стего (заполненный контейнер) S:

$$S = F(T,\ M_S,\ K_2);$$

$$K_2 = \{r \pm k_1,\ g \pm k_2,\ b \pm k_3\},$$

где r, g, b — исходные цветовые координаты символов;

– извлечение сообщения M:

$$M = F(S,\ K_2,\ K_1).$$

Нашей задачей в дальнейшем исследовании является определение вида F и влияния T, M_S, K_2 (параметров функции) на конечное значение S.

Заключение. В статье приводится формальное описание математической модели процесса формирования секретного сообщения, его осаждения в текст-контейнер и извлечения на основе стеганографических преобразований. Скрываемая информация размещается в цветовых координатах символов текста. Проведена аналогия между битовой картой изображения и символом текста. Основу математической модели составляет описание процесса создания цифрового водяного знака.

Литература

1. Bennett K. Linguistic steganography: Survey, analysis, and robustness concerns for hiding information in text. West Lafayette: Purdue Univ., 2004. 30 p.

2. Конахович Г. Ф., Пузыренко А. Ю. Компьютерная стеганография. Теория и практика. Киев: МК-Пресс, 2006. 288 с.

3. Пласковицкий В. А., Шутько Н. П. Управление защитой информации на основе стеганографических методов // Автоматический контроль и автоматизация производственных процессов: междунар. науч.-техн. конф., Минск, апрель 2012 г. / Белорус. гос. технол. ун-т. Минск: БГТУ, 2012. С. 283–285.

4. Urbanovich N. Development, analysis of efficiency and performance in an electronic textbook methods of text steganography // Printing future days: 4th International Scientific Conference on Printing and Media Technology, Germany, Chemnitz, november 2011. Chemnitz, 2011. P. 189–193.

5. Урбанович Н. П. Исследование эффективности стеганографических методов скрытия информации в тексте // Новые математические методы и компьютерные технологии в проектировании, производстве и научных исследованиях: XIII респ. науч. конф. студентов и аспирантов, Гомель, 21–23 марта 2011 г.: в 2 ч. / Гомел. гос. ун-т им. Ф. Скорины. Гомель, 2011. Ч. 2. С. 27–28.

6. Urbanovich N., Plaskovitsky V. The use of steganographic techniques for protection of intellectual property rights // New Electrical and Electronic Technologies and their Industrial Implementation: 7-th Int. Konf., Zakopane, Poland, june 2011. Zakopane, 2011. P. 147–148.

7. Shutko N. Text steganography as an effective instrument of protection of the copyright on electronic document // New Electrical and Electronic Technologies and their Industrial Implementation: 8-th Int. Konf., Zakopane, Poland, june 18–21, 2013. Zakopane, 2013. P. 147.

8. Филиппович А. Ю. Компьютерные шрифты: форматы, кодировка, растеризация. М.: МГУП, 2012. 12 с.

9. Свидетельство о регистрации компьютерной программы Sword v.1.0 / В. А. Пласковицкий, Н. П. Шутько. № 383 от 04.01.2012 // Реестр Нац. центра интеллектуал. собственности Респ. Беларусь. Минск, 2012.

10. Document Marking and Identification using Both Line and Word Shifting / S. H. Low [et al.]. Boston: Infocom, 1995. 8 p.

Поступила 20.03.2014